DOCUMENT RESUME

ED 406 393 TM 026 256

AUTHOR	Blankmeyer, Eric
TITLE	High-Breakdown Regression by Least Quartile Differences.
PUB DATE	96
NOTE	16p.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Computer Programs (101)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computer Software; *Estimation (Mathematics); Least Squares Statistics; *Regression (Statistics); Research Methodology; *Robustness (Statistics)
IDENTIFIERS	Outliers

ABSTRACT

A high-breakdown estimator is a robust statistic that can withstand a large amount of contaminated data. In linear regression, high-breakdown estimators can detect outliers and distinguish between good and bad leverage points. This paper summarizes the case for high-breakdown regression and emphasizes the least quartile difference estimator (LQD) proposed by C. Croux, P. J. Rousseeuw, and O. Hossjer (1994). This regression method examines the absolute differences between every pair of residuals and minimizes the first quartile of these differences with an adjustment for degrees of freedom. LQD is affine equivalent and has a 50% breakdown point, the highest possible. Its asymptotic efficiency is about 67% of least-squares, so LQD should be able to deal with anomalous observations and should also perform well when the data are not contaminated. Although interest in the approach is growing, software is still not widely available. An appendix presents a BASIC computer program for one type of high-breakdown regression. (Contains 17 references.) (SLD)

High-Breakdown Regression by Least Quartile Differences

Eric Blankmeyer
Department of Finance and Economics
Southwest Texas State University
San Marcos, TX 78666
e-mail: eb01@business.swt.edu
telephone 512-245-3253

Abstract. This paper is a concise presentation of high-breakdown regression with emphasis on the least quartile difference (LQD) estimator proposed by Croux, Rousseeuw, and Hossjer. The robustness of high-breakdown regression is discussed, and a Basic program for LQD is provided.

Key words. Breakdown point, least median of squares, least quartile difference, linear regression, robust statistics.

High-breakdown Regression by Least Quartile Differences

Introduction

A high-breakdown estimator is a robust statistic which can withstand a large amount of contaminated data. In linear regression, high-breakdown estimators can detect outliers and distinguish between good and bad leverage points. Although most high-breakdown estimators are based on resampling schemes which require extensive computation, recent improvements in hardware and algorithms have made these robust methods practical.

This paper summarizes the case for high-breakdown regression. Although interest in the topic is growing and the literature is extensive, software is still not widely available. Some public-domain sources are mentioned, and an appendix contains a Basic program for one type of high-breakdown regression.

The breakdown point

If a few bad observations can induce an arbitrarily large bias in an estimator, the estimator is said to have a low breakdown point (Rousseeuw and Leroy 1987, chapter 1). For instance, in a sample of n data, the breakdown point of a least-squares (LS) regression is 1/n. Just one outlier, if it is bad enough, can throw the LS line indefinitely far off target. On the other hand, a robust method has a high breakdown point; it can deal with a lot of contamination. The highest achievable breakdown point is fifty percent. If more than half the observations are contaminated, a linear regression method cannot distinguish the good observations from the bad.

One kind of contamination is a regression outlier, a stray observation on the dependent variable. Another kind of contamination is a bad leverage point, a stray observation on an independent variable. A high-breakdown method is unaffected by regression outliers and bad leverage points but does not ignore good leverage points, those observations which lie apart from the other data but are still close to the regression line. Good leverage points improve the precision of the estimates.

In conjunction with other criteria, the breakdown concept can be used to evaluate the sturdiness of various statistics. The sample median is very robust; it attains the maximum breakdown point. The L1 norm, a multivariate version of the median, minimizes the sum of absolute residuals rather than the sum of squared residuals (Dodge 1987, 1992). The L1 norm is not vulnerable to regression outliers but is vulnerable to bad leverage points, so it shares the low breakdown point of LS, as do the M-estimators of Huber (1981). The generalized M-estimators (Hampel et al. 1986, chapter 6) can cope with bad leverage points. Unfortunately, their breakdown point is inversely related to the number of variables in the model (Rousseeuw and Leroy 1987, chapter 6), so they may be affected by aberrant observations in high-dimensional regression problems.

The LS residuals may not be reliable indicators of outliers. If the regression line has broken down, some good observations will have large LS residuals and some bad observations will have small residuals. The standard tests for heteroscedasticity based on the LS residuals can be quite misleading. The same remark applies to classical diagnostics like Cook's distance and the "hat" matrix. To end up with a robust estimate, one has to start with a robust estimate.

It is tempting to seek robustness by examining the variables one at a time before running the regression. Robust measures of location and dispersion can be computed, while a histogram of each variable may suggest a simple transformation to reduce skewness. This common-sense approach, involving a careful inspection of the data, is certainly worthwhile. However, it is not guaranteed to reveal all the outliers and bad leverage points because the troublesome data are aberrant with respect to the hypothesized regression line, not with respect to a particular variable.

Of course, some robustness problems are more tractable than others. For example, certain models are not subject to bad leverage points. This is trivially the case for univariate samples, but it is also true when the independent variables in a regression model are deterministic. Examples are simple trends and dummy variables, including ANOVA models. The L1 norm or an M-estimator is very appropriate in these cases, where the possibility of contamination is limited to the dependent variable.

In summary, a high-breakdown point is a desirable robustness property, especially in regression and other multivariate models where leverage points can occur. As it happens, combining high breakdown with other good statistical properties is not always straightforward. In particular, a linear regression procedure should be reasonably efficient compared to LS when the data are free of contamination. Moreover, the procedure ought to behave well under linear transformations of the variables: the regression coefficients should adjust in an obvious way when the dependent variable or an independent variable undergoes a shift of origin or a change of scale. LS has this affine equivariance, of course, and so should robust regression.

High-breakdown regression

Rousseeuw proposed the first high-breakdown, affine-equivariant estimator for linear regression (Rousseeuw 1984, Rousseeuw and Leroy 1987, chapters 1-3, 5). It is called least median of squares (LMS) because the regression line minimizes the median squared residual. The geometric interpretation for bivariate regression is instructive. LMS finds the narrowest strip that covers at least half the observations. (Narrowness is measured in the direction of the dependent variable.) The regression line lies in the middle of the strip. In other words, LMS tries to fit the majority of the observations, ignoring the aberrant data. This explains its asymptotic breakdown point of fifty percent. .

When the regression errors are in fact drawn from a well-behaved gaussian distribution, LMS suffers from the low efficiency common to many robust techniques. There are several proposals for high-breakdown, efficient regression estimators. One can use a two-step procedure, in which the LMS residuals are used as weights in a LS regression. Various weighting schemes have been discussed (Rousseeuw and Leroy 1987, chapter 3).

Alternatively, the median squared residual can be replaced by a more efficient minimand. Croux, Rousseeuw, and Hossjer (1994) introduced the Least Quartile Distance (LQD) estimator. This regression method examines the absolute differences between every pair of residuals and minimizes the first quartile of those differences (with an adjustment for degrees of freedom). It is therefore a multivariate version of the mode, a "nearest neighbor" estimator which is unaffected by skewed data. LQD is affine equivariant and has a fifty percent breakdown point, the highest possible. Its asymptotic efficiency is about 67 percent of LS --very high for a robust method. Therefore, LQD should be able to deal with anomalous observations and at the same time should perform well when the data are not contaminated.

Like LMS, LQD is computed using a resampling scheme. To estimate k regression coefficients, a subsample of k observations is drawn at random from the data set. The regression coefficients are calculated; and the residuals are computed for the entire sample. The differences between each pair of residuals are obtained, and the first quartile of their absolute values is evaluated. This process is repeated many times, and the final regression line is based on the subsample with the least quartile difference.

Rousseeuw and Leroy (1987, chapter 5) have discussed the amount of resampling needed to obtain a robust estimate with high probability. Some guidelines are offered in the Basic program itself. [Rousseeuw (1993) has recently proposed an alternative sampling scheme which has not, however, been implemented in the Basic program for LQD.] Obviously, LQD regression is much more "computer intensive" than LS or even the L1 norm. (The latter is a linear programming problem.) For n observations, the LQD requires, in every subsample, an order statistic of n(n-1)/2 differences between pairs of residuals. Even for moderate n, the calculations would be impractical were it not for Rousseeuw and Croux's clever algorithm.

When the LQD regression has flagged several observations as potential outliers, the researcher will probably want to scrutinize those data. They will have to be discarded, downweighted, or validated and reinstated. Then the researcher may choose to apply LS to the revised sample in order to examine the usual F-statistic, t-ratios, and so on. Since the sample has undergone a complicated "pretest," those diagnostics are no longer strictly valid for classical hypothesis tests. However, they may provide a general impression of the model's adequacy.

Nonlinear parameters pose further computational and conceptual issues for high-breakdown regression, some of which have been explored by Stromberg (1993) and Stromberg and Ruppert (1992).

Programs for high-breakdown regression

Appendix A contains a Basic program for LQD. The program is simple but serviceable, with few bells and whistles. Users are encouraged to add code to enhance the flexibility of input, the labeling of output, the treatment of missing data, graphics, and other features.

The on-line statistical resource STATLIB currently provides Fortran source code for several procedures. This author is aware of Rousseeuw and Leroy's PROGRESS program for LMS regression and their MINVOL program for covariance matrices; Rousseeuw and Croux's Sn and Qn scale estimators; and Douglas Hawkins' feasible-solution algorithms (FSA) for regression and covariance matrices. Rocke and Woodruff have provided C code for a program to estimate high-breakdown covariance matrices. Other public-domain and commercial implementations are no doubt available.

The literature on robust regression includes many well-studied data sets, both real and hypothetical, which allow researchers to compare LS, L1, LMS, LQD and other methods. In particular, the monograph by Rousseeuw and Leroy (1987) contains numerous examples.

References

Bassett, Jr., Gilbert and Roger Koenker (1978). "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, Vol. 73, pp. 618-622.

Croux, Christophe, Peter J. Rousseeuw. and Ola Hossjer (1994). "Generalized S-Estimators," *Journal of the American Statistical Association*, Vol. 89, No. 428, pp. 1271-1281.

Dodge, Yadolah (editor) (1987). *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Amsterdam: North-Holland.

Dodge, Yadolah (editor) (1992). L1-Statistical Analysis and Related Methods. Amsterdam: North-Holland.

Hadi, Ali S. and Jeffrey S. Simonoff (1993). "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, Vol. 88, No. 424, pp. 1264-1272.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: the Approach Based on Influence Functions*. New York: Wiley.

Hawkins, Douglas M. (1993). "The Accuracy of Elemental Set Approximations for Regressions," *Journal of the American Statistical Association*, Vol. 88, No. 422, pp. 580-589.

Hettmansperger, Thomas P. and Simon J. Sheather (1992). "A Cautionary Note on the Method of Least Median Squares," *The American Statistician*, Vol. 46, No. 2, pp. 79-83.

Huber, Peter J. (1981). *Robust Statistics*. New York: Wiley.

Rocke, David M. and David L. Woodruff (1996). "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, Vol. 91, No. 435, pp. 1047-1061.

Rousseeuw, Peter J. (1984). "Least Median of Squares Regression," *Journal of the American Statistical Association*, Vol. 79, No. 388, pp. 871-880.

Rousseeuw, Peter J. (1993). "A resampling design for computing high-breakdown regression," *Statistics and Probability Letters*, Vol. 18, No. 2, pp. 125-128.

Rousseeuw, Peter J. and Annick M. Leroy (1987). *Robust Regression and Outlier Detection*. New York, NY: Wiley.

Rousseeuw, Peter J. and Bert C. van Zomeren (1990). "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, Vol. 85, No. 411, pp. 633-651 (with discussion).

Rousseeuw, Peter J. and Christophe Croux (1993). "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, Vol. 88, No. 424, pp. 1273-1283.

Stromberg, Arnold J. (1993). "Computation of High Breakdown Nonlinear Regression Parameters," *Journal of the American Statistical Association*, Vol. 88, No. 421, pp. 237-244.

Stromberg, Arnold J. and David Ruppert (1992). "Breakdown in Nonlinear Regression," *Journal of the American Statistical Association*, Vol. 87, No. 420, pp. 991-997.

Appendix A. A Basic Program for the Least Quartile Difference Regression

```
REM  THIS PROGRAM COMPUTES A ROBUST LINEAR REGRESSION CALLED
REM  THE LEAST QUARTILE DIFFERENCE ESTIMATOR (LQD). IT WAS
REM  PROPOSED IN THE FOLLOWING PAPER:
REM
REM     CHRISTOPHE CROUX, PETER J. ROUSSEEUW, AND OLA
REM     HOSSJER (1994), "GENERALIZED S-ESTIMATORS,"
REM     JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION,
REM     89, 1271-1281.
REM
REM  THE LQD IS CLOSELY RELATED TO OTHER HIGH-BREAKDOWN,
REM  COMPUTER-INTENSIVE REGRESSION METHODS LIKE LEAST
REM  MEDIAN OF SQUARES AND LEAST TRIMMED SQUARES. WHEN THE
REM  RANDOM ERRORS ARE IN FACT GAUSSIAN, THE LQD IS MORE
REM  EFFICIENT THAN THESE OTHER HIGH-BREAKDOWN METHODS.
REM  (ITS ASYMPTOTIC EFFICIENCY IS ABOUT 67 PERCENT
REM  COMPARED TO LEAST SQUARES AT A GAUSSIAN DISTRIBUTION.)
REM  LQD RESEMBLES A MODE; IT IS A "NEAREST NEIGHBOR"
REM  ESTIMATOR WHICH COMPARES THE ABSOLUTE DIFFERENCE
REM  BETWEEN EVERY PAIR OF RESIDUALS AND MINIMIZES A CERTAIN
REM  ORDER STATISTIC OF THOSE DIFFERENCES.
REM
REM  THE CURRENT VERSION OF THE PROGRAM IS DIMENSIONED FOR
REM  7 INDEPENDENT VARIABLES AND 500 OBSERVATIONS. A CONSTANT
REM  IS AUTOMATICALLY INCLUDED IN THE REGRESSION EQUATION.
REM  THE PROGRAM ASSUMES THAT THE ROWS OF THE DATA MATRIX
REM  ARE OBSERVATIONS, WHILE THE COLUMNS ARE VARIABLES. THE
REM  LAST COLUMN SHOULD CONTAIN THE DEPENDENT VARIABLE. TO
REM  MAINTAIN NUMERICAL ACCURACY, THE USER SHOULD SCALE THE
REM  DATA SO THAT ALL THE VARIABLES HAVE THE SAME ORDER OF
REM  MAGNITUDE.
REM
REM  LIKE MOST OTHER HIGH-BREAKDOWN REGRESSION METHODS,
REM  LQD IS BASED ON A RESAMPLING SCHEME; THE REGRESSION
REM  LINE IS FIT TO A LARGE NUMBER OF SUBSAMPLES, AND THE
REM  LINE THAT MINIMIZES A ROBUST CRITERION IS SELECTED. USE
REM  ENOUGH SUBSAMPLES (ITR) TO PROVIDE VIRTUAL ASSURANCE
REM  THAT THERE WILL BE SEVERAL UNCONTAMINATED SUBSAMPLES.
REM  IN THEIR MONOGRAPH, ROBUST REGRESSION AND OUTLIER
REM  DETECTION (NEW YORK: WILEY, 1987), P. J. ROUSSEEUW AND
REM  A. M. LEROY DISCUSS THE RESAMPLING ISSUE IN DETAIL AND
REM  OFFER THE FOLLOWING ROUGH GUIDELINES:
REM
REM         INDEPENDENT                  MINIMUM NUMBER
REM          VARIABLES                   OF SUBSAMPLES
REM              1                             500
REM              2                           1,000
REM              3                           1,500
REM              4                           2,000
REM              5                           2,500
REM              6 OR MORE                   3,000
REM
REM  FOR FURTHER DISCUSSION, THE USER SHOULD CONSULT THE
REM  REFERENCES CITED ABOVE.
REM
REM  THE GAUSS-JORDAN ROUTINE FOR MATRIX INVERSION IS
REM  ADAPTED FROM W. W. COOLEY AND P. R. LOHNES (1962),
REM  MULTIVARIATE PROCEDURES FOR THE SOCIAL SCIENCES,
REM  NEW YORK: WILEY.
REM  THE FUNCTIONS AND SUBROUTINES IN THIS PROGRAM ARE
REM  ADAPTED FROM "QN.FOR" -- A FORTRAN PROGRAM FOR
REM  ROBUST SCALE ESTIMATES WRITTEN BY P. J. ROUSSEEUW
REM  AND HIS COLLEAGUES.
REM  DETAILS ON THE QN SCALE ESTIMATE ARE GIVEN IN A
REM  PAPER BY P. J. ROUSSEEUW AND C. CROUX (1993):
REM    "ALTERNATIVES TO THE MEDIAN ABSOLUTE DEVIATION,"
REM    JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION,
REM    88, 1273-1283.
REM
REM  THERE IS NO WARRANTY, EXPRESSED OR IMPLIED, FOR THIS
REM  PROGRAM. ITS SUITABILITY FOR COMMERCIAL USE OR FOR
REM  ANY PARTICULAR PURPOSE IS NOT GUARANTEED.
REM
4002 DEFINT I-N
4004 DEFDBL C-D
4006 DECLARE FUNCTION SCALE (X(), N,K)
4008 DECLARE FUNCTION FMED (A(), N, NH)
4010 DECLARE FUNCTION WHIMED (A(), IW(), NX)
4012 DECLARE SUB SORT (A(), N, B())
4015 RANDOMIZE TIMER
4020 PRINT "WHAT IS THE INPUT FILE ?"
4025 INPUT INFILE$
4030 PRINT "WHAT IS THE OUTPUT FILE ?"
4035 INPUT OUTFILE$
4040 OPEN INFILE$ FOR INPUT AS #1
4045 OPEN OUTFILE$ FOR OUTPUT AS #2
4050 PRINT "HOW MANY OBSERVATIONS ?"
4055 INPUT N
4060 REM    THE DEPENDENT VARIABLE MUST BE IN THE LAST COLUMN
4065 REM    OF THE DATA ARRAY.
4070 PRINT "HOW MANY VARIABLES (INDEPENDENT + DEPENDENT) ?"
4075 INPUT K
4080 PRINT "HOW MANY SUBSAMPLES SHOULD BE DRAWN ?"
4085 INPUT ITR
4090 DIM VI(500, 8), R(500), BLQD(8), C(8, 8), D(8)
4095 DIM VD(500), IPIVOT(8), PIVOT(8), INDEX(8, 2), M(8)
4100 CRITMIN = 1000000#
4110 PIVMIN = .000001
4120 RK = K
4130 RN = N
4135 REM   READ THE DATA
4140 FOR I = 1 TO N
4150 FOR J = 1 TO K
4160 INPUT #1, VI(I, J)
4170 NEXT J
4180 NEXT I
4185 REM   INCLUDE A CONSTANT (INTERCEPT) IN THE REGRESSION
```

```
4190 FOR I = 1 TO N
4200 VD(I) = VI(I, K)
4210 VI(I, K) = 1!
4220 NEXT I
4235 REM   START ITERATIONS; CHOOSE A RANDOM SUBSAMPLE OF K DATA
4240 FOR L = 1 TO ITR
4250 PRINT "INTERATION  "; L
4260 FOR I = 1 TO K
4262 M(I) = INT(RND * N) + 1
4264 NEXT I
4266 FOR I = 1 TO K
4268 FOR J = 1 TO K
4270 IF I = J GOTO 4274
4272 IF M(I) = M(J) GOTO 4260
4274 NEXT J
4276 NEXT I
4278 FOR I = 1 TO K
4280 MI = M(I)
4282 D(I) = VD(MI)
4290 FOR J = 1 TO K
4300 C(I, J) = VI(MI, J)
4310 NEXT J
4320 NEXT I
4325 REM   INVERT THE KxK MATRIX
4330 DETERM = 1!
4340 FOR J = 1 TO K
4350 IPIVOT(J) = 0
4360 NEXT J
4370 FOR I = 1 TO K
4380 CMAX = 0!
4390 FOR J = 1 TO K
4400 IF IPIVOT(J) = 1 GOTO 4480
4410 FOR JJ = 1 TO K
4420 IF IPIVOT(JJ) = 1 GOTO 4470
4422 IF IPIVOT(JJ) > 1 GOTO 4890
4430 IF ABS(CMAX) >= ABS(C(J, K)) GOTO 4470
4440 IROW = J
4450 ICOLUM = JJ
4460 CMAX = C(J, JJ)
4470 NEXT JJ
4480 NEXT J
4485 IPIVOT(ICOLUM) = IPIVOT(ICOLUM) + 1
4490 IF IROW = ICOLUM GOTO 4590
4500 DETERM = -DETERM
4510 FOR J = 1 TO K
4520 ZWAP = C(IROW, J)
4530 C(IROW, J) = C(ICOLUM, J)
4540 C(ICOLUM, J) = ZWAP
4550 NEXT J
4560 ZWAP = D(IROW)
4570 D(IROW) = D(ICOLUM)
4580 D(ICOLUM) = ZWAP
4590 INDEX(I, 1) = IROW
4600 INDEX(I, 2) = ICOLUM
```

```
4610 PIVOT(I) = C(ICOLUM, ICOLUM)
REM  IF THE SUBSAMPLE IS SINGULAR, OR NEARLY SO, DISCARD
REM  IT AND DRAW ANOTHER SUBSAMPLE.
4612 IF ABS(PIVOT(I)) < PIVMIN GOTO 4260
4620 DETERM = DETERM * PIVOT(I)
4630 C(ICOLUM, ICOLUM) = 1!
4640 FOR J = 1 TO K
4650 C(ICOLUM, J) = C(ICOLUM, J) / PIVOT(I)
4660 NEXT J
4670 D(ICOLUM) = D(ICOLUM) / PIVOT(I)
4680 FOR JJ = 1 TO K
4690 IF JJ = ICOLUM GOTO 4760
4700 T = C(JJ, ICOLUM)
4710 C(JJ, ICOLUM) = 0!
4720 FOR J = 1 TO K
4730 C(JJ, J) = C(JJ, J) - C(ICOLUM, J) * T
4740 NEXT J
4750 D(JJ) = D(JJ) - D(ICOLUM) * T
4760 NEXT JJ
4770 NEXT I
4780 FOR I = 1 TO K
4790 JJ = K + 1 - I
4800 IF INDEX(JJ, 1) = INDEX(JJ, 2) GOTO 4880
4810 JROW = INDEX(JJ, 1)
4820 JCOLUM = INDEX(JJ, 2)
4830 FOR J = 1 TO K
4840 ZWAP = C(J, JROW)
4850 C(J, JROW) = C(J, JCOLUM)
4860 C(K, JCOLUM) = ZWAP
4870 NEXT J
4880 NEXT I
4890 REM    COMPUTE ROBUST RESIDUALS FOR THE ENTIRE SAMPLE
4900 FOR I = 1 TO N
4910 SUM = 0!
4920 FOR J = 1 TO K
4930 SUM = SUM + VI(I, J) * D(J)
4940 NEXT J
4950 R(I) = VD(I) - SUM
4970 NEXT I
REM  COMPUTE THE ROBUST CRITERION QN.
5210 CRIT = SCALE(R(), N, K)
5220 IF CRITMIN <= CRIT GOTO 5300
5230 CRITMIN = CRIT
5270 FOR J = 1 TO K
5280 BLQD(J) = D(J)
5290 NEXT J
5300 NEXT L
5310 PRINT #2, "ROBUST SCALE ESTIMATE  ", CRITMIN
5435 REM   PRINT THE LQD REGRESSION COEFFICIENTS
5440 PRINT #2, "LQD REGRESSION COEFFICIENTS (CONSTANT LAST)"
5450 FOR J = 1 TO K
5460 PRINT #2, J, BLQD(J)
5470 NEXT J
REM  STANDARDIZE THE ROBUST RESIDUALS AND IDENTIFY POTENTIAL
REM  OUTLIERS.
5475 PRINT #2, "   "
5477 PRINT #2, "STANDARDIZED ROBUST RESIDUALS >= 2.5"
5480 FOR I = 1 TO N
5490 SUM = 0!
5500 FOR J = 1 TO K
5510 SUM = SUM + VI(I, J) * BLQD(J)
5520 NEXT J
5530 R(I) = (VD(I) - SUM)/CRITMIN
5540 IF ABS(R(I)) < 2.5 THEN GOTO 5560
5550 PRINT #2, I, R(I)
5560 NEXT I
5570 END

940 FUNCTION FMED (A(), N, NH)
945 DEFINT I-N
950 DIM B(500)
955 FOR I = 1 TO N
960 B(I) = A(I)
965 NEXT I
980 LL = 1
990 LR = N
1000 IF LL >= LR GOTO 1210
1010 AX = B(NH)
1020 JNC = LL
1030 J = LR
1040 IF JNC > J GOTO 1180
1050 IF B(JNC) >= AX GOTO 1080
1060 JNC = JNC + 1
1070 GOTO 1050
1080 IF B(J) <= AX GOTO 1110
1090 J = J - 1
1100 GOTO 1080
1110 IF JNC > J GOTO 1170
1120 WA = B(JNC)
1130 B(JNC) = B(J)
1140 B(J) = WA
1150 JNC = JNC + 1
1160 J = J - 1
1170 GOTO 1040
1180 IF J < NH THEN LL = JNC
1190 IF NH < JNC THEN LR = J
1200 GOTO 1000
1210 FMED = B(NH)
1220 END FUNCTION

1500 FUNCTION SCALE (X(), N,K)
1502 DEFINT I-N
1503 DIM Y(500), WORK(500), LEFT(500)
1504 DIM IRIGHT(500), IWEIGHT(500), IQ(500), IP(500)
1506 IH = (N + K + 1)\2
1508 KK = IH * (IH - 1) \ 2
1510 CALL SORT(X(), N, Y())
1512 FOR I = 1 TO N
```

```
1514 LEFT(I) = N - I + 2
1516 IRIGHT(I) = N
1518 NEXT I
1520 JHELP = N * (N + 1) \ 2
1522 KNEW = KK + JHELP
1524 NL = JHELP
1526 NR = N * N
1628 IFOUND = 0
1630 REM
1632 IF (NR - NL > N) AND (IFOUND = 0) THEN
1634 J = 1
1636 FOR I = 2 TO N
1638 IF LEFT(I) <= IRIGHT(I) THEN
1640 IWEIGHT(J) = IRIGHT(I) - LEFT(I) + 1
1642 JHELP = LEFT(I) + IWEIGHT(J) \ 2
1644 WORK(J) = Y(I) - Y(N + 1 - JHELP)
1646 J = J + 1
1648 END IF
1650 NEXT I
1652 TRIAL = WHIMED(WORK(), IWEIGHT(), J - 1)
1654 J = 0
1656 FOR I = N TO 1 STEP -1
1658 REM
1660 IF J < N AND ((Y(I) - Y(N - J)) < TRIAL) THEN
1662 J = J + 1
1664 GOTO 1658
1666 END IF
1668 IP(I) = J
1670 NEXT I
1672 J = N + 1
1674 FOR I = 1 TO N
1676 REM
1678 IF ((Y(I) - Y(N - J + 2)) > TRIAL) THEN
1680 J = J - 1
1682 GOTO 1676
1684 END IF
1686 IQ(I) = J
1688 NEXT I
1690 ISUMP = 0
1692 ISUMQ = 0
1694 FOR I = 1 TO N
1696 ISUMP = ISUMP + IP(I)
1698 ISUMQ = ISUMQ + IQ(I) - 1
1700 NEXT I
1701 IF KNEW <= ISUMP THEN
1702 FOR I = 1 TO N
1704 IRIGHT(I) = IP(I)
1706 NEXT I
1708 NR = ISUMP
1710 ELSE
1712 IF KNEW > ISUMQ THEN
1714 FOR I = 1 TO N
1716 LEFT(I) = IQ(I)
1718 NEXT I
```

```
1720 NL = ISUMQ
1722 ELSE
1724 QN = TRIAL
1726 IFOUND = -1
1728 END IF
1730 END IF
1732 GOTO 1630
1734 END IF
1736 IF IFOUND = 0 THEN
1738 J = 1
1740 FOR I = 2 TO N
1742 IF LEFT(I) <= IRIGHT(I) THEN
1744 FOR JJ = LEFT(I) TO IRIGHT(I)
1746 WORK(J) = Y(I) - Y(N - JJ + 1)
1748 J = J + 1
1750 NEXT JJ
1752 END IF
1754 NEXT I
1756 QN = FMED(WORK(), J - 1, KNEW - NL)
1758 END IF
1760 IF N <= 9 THEN
1762 IF N = 2 THEN DN = .399
1764 IF N = 3 THEN DN = .994
1766 IF N = 4 THEN DN = .512
1768 IF N = 5 THEN DN = .844
1770 IF N = 6 THEN DN = .611
1772 IF N = 7 THEN DN = .857
1774 IF N = 8 THEN DN = .669
1776 IF N = 9 THEN DN = .872
1778 ELSE
1780 IF (N MOD 2) = 1 THEN DN = N / (N + 1.4)
1782 IF (N MOD 2) = 0 THEN DN = N / (N + 3.8)
1784 END IF
1786 SCALE = DN * 2.2219 * QN
1788 END FUNCTION

570 SUB SORT (A(), N, B())
575 DEFINT I-N
580 DIM JLV(500), JRV(500)
585 FOR I = 1 TO N
590 B(I) = A(I)
595 NEXT I
600 JSS = 1
610 JLV(1) = 1
620 JRV(1) = N
630 JNDL = JLV(JSS)
640 JR = JRV(JSS)
650 JSS = JSS - 1
660 JNC = JNDL
670 J = JR
680 JTWE = (JNDL + JR) \ 2
690 XX = B(JTWE)
700 IF B(JNC) >= XX GOTO 730
710 JNC = JNC + 1
```

```
720 GOTO 700
730 IF XX >= B(J) GOTO 760
740 J = J - 1
750 GOTO 730
760 IF JNC > J GOTO 806
770 AMM = B(JNC)
780 B(JNC) = B(J)
790 B(J) = AMM
800 JNC = JNC + 1
804 J = J - 1
806 IF JNC <= J GOTO 700
808 IF (J - JNDL) < (JR - JNC) GOTO 822
810 IF JNDL >= J GOTO 818
812 JSS = JSS + 1
814 JLV(JSS) = JNDL
816 JRV(JSS) = J
818 JNDL = JNC
820 GOTO 832
822 IF JNC >= JR GOTO 830
824 JSS = JSS + 1
826 JLV(JSS) = JNC
828 JRV(JSS) = JR
830 JR = J
832 IF JNDL < JR GOTO 660
834 IF JSS <> 0 GOTO 630
836 END SUB

1900 FUNCTION WHIMED (A(), IW(), NX)
1902 DEFINT I-N
1904 DIM ACAND(500), IWCAND(500)
1906 NN = NX
1908 IWTOTAL = 0
1910 FOR I = 1 TO NN
1912 IWTOTAL = IWTOTAL + IW(I)
1914 NEXT I
1916 IWREST = 0
1918 REM
1920 TRIAL = FMED(A(), NN, NN \ 2 + 1)
1922 IWLEFT = 0
1924 IWMID = 0
1926 IWRIGHT = 0
1928 FOR I = 1 TO NN
1930 IF A(I) < TRIAL THEN
1932 IWLEFT = IWLEFT + IW(I)
1934 ELSE
1936 IF A(I) > TRIAL THEN
1938 IWRIGHT = IWRIGHT + IW(I)
1940 ELSE
1942 IWMID = IWMID + IW(I)
1944 END IF
1946 END IF
1948 NEXT I
1950 IF ((2 * IWREST + 2 * IWLEFT) > IWTOTAL) THEN
1952 KCAND = 0
```

```
1954 FOR I = 1 TO NN
1956 IF A(I) < TRIAL THEN
1958 KCAND = KCAND + 1
1960 ACAND(KCAND) = A(I)
1962 IWCAND(KCAND) = IW(I)
1964 END IF
1966 NEXT I
1968 NN = KCAND
1970 ELSE
1972 IF ((2 * IWREST + 2 * IWLEFT + 2 * IWMID) > IWTOTAL) THEN
1974 WHIMED = TRIAL
1976 GOTO 2014
1978 ELSE
1980 KCAND = 0
1982 FOR I = 1 TO NN
1984 IF A(I) > TRIAL THEN
1986 KCAND = KCAND + 1
1988 ACAND(KCAND) = A(I)
1990 IWCAND(KCAND) = IW(I)
1992 END IF
1994 NEXT I
1996 NN = KCAND
1998 IWREST = IWREST + IWLEFT + IWMID
2000 END IF
2002 END IF
2004 FOR I = 1 TO NN
2006 A(I) = ACAND(I)
2008 IW(I) = IWCAND(I)
2010 NEXT I
2012 GOTO 1918
2014 REM
2016 END FUNCTION
```

TM026256

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: High-Breakdown Regression by Least Quartile Differences	
Author(s): ERIC BLANKMEYER	
Corporate Source: SOUTHWEST TEXAS STATE UNIV.	Publication Date: AUG. 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Level 1

Sample sticker to be affixed to document □

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Eric Blankmeyer	Position: PROFESSOR
Printed Name: ERIC BLANKMEYER	Organization: SOUTHWEST TEXAS STATE U.
Address: DEPT. FINANCE / ECONOMICS SOUTHWEST TEXAS STATE U. SAN MARCOS TX 78666	Telephone Number: (512) 245-3253
	Date: NOV. 28 1996

OVER

BEST COPY AVAILABLE